Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.6, No.3, 2015

A Survey of Model Used for Web User's Browsing Behavior Prediction

Pradipsinh K. Chavda
Department of CS&E,Governmernt Engineering Collage,Modasa Aravalli,Gujarat,India
E-mail: pradipchavda.it@gmail.com

Prof. Jitendra S.Dhobi
Department of CS&E,Governmernt Engineering Collage,Modasa Aravalli,Gujarat,India
E-mail: jsdhobi@gmail.com

Abstract:
The motivation behind the work is that the prediction of web user's browsing behavior while serving the Internet, reduces the user's browsing access time and avoids the visit of unnecessary pages to ease network traffic. Various models such as fuzzy interference models, support vector machines (SVMs), artificial neural networks (ANNs), association rule mining (ARM), k-nearest neighbor(kNN) Markov model, Kth order Markov model, all-Kth Markov model and modified Markov model were proposed to handle Web page prediction problem. Many times, the combination of two or more models were used to achieve higher prediction accuracy. This research work introduces the Support Vector Machines for web page prediction. The advantages of using support vector machines is that it offers most robust and accurate classification due to their generalized properties with its solid theoretical foundation and proven effectiveness. Web contains enormous amount of data and web data increases exponentially but the training time for Support vector machine is very large. That is, SVM's suffer from a widely recognized scalability problem in both memory requirement and computation time when the input dataset is too large. To address this, I aimed at training the Support vector machine model in MapReduce programming model of Hadoop framework, since the MapReduce programming model has the ability to rapidly process large amount of data in parallel. MapReduce works in tandem with Hadoop Distributed File System (HDFS). So proposed approach will solve the scalability problem of present SVM algorithm.
Keywords: Web Page Prediction, Support Vector Machines, Hadoop, MapReduce, HDFS.

1. Introduction

In order to predict web-browsing behavior, we need the knowledge of web data management and mining techniques. Enormous amount of data on web has to be managed and mined so as to extract the relevant information. Web data management involves the appropriate data representation scheme. Web database management include query processing, meta-data management, security and integrity. Web data mining is important thing to be known since it is the key aspect of web information management. Web data mining involves three types of mining web data namely, mining data sources, mining web structure and mining web usage patterns. Mining data sources helps to extract patterns from data sources. Mining web structure helps to improve search engines. Mining web usage patterns helps to give advice to users while browsing [1]. Web Usage Mining is the automatic discovery of user access patterns from Web servers. Organizations collect large volumes of data in their daily operations, generated automatically by Web servers that are collected in Web access log files. Analysis of these access data can provide useful information for server performance enhancements, restructuring a Web site, and direct marketing in e-commerce [2].

Web page prediction [1] is a classification problem which helps to forecast the next set of Web pages that a user may visit based on the knowledge of the previously visited pages. This requires the understanding of mining the web usage patterns. The web usage mining involves discovery and analysis of usage patterns from web logs.

In Web page prediction, the available source of training data is the users' sessions. The user sessions give the information regarding the click stream data i.e. the path viewed by the users and their access time on each page. The information regarding the path visited, browsing rate and relative duration of access time is considered while discovering users' interest on web.

In Web page prediction, both preprocessing and prediction challenges have to be addressed. Preprocessing challenges are, handling large amount of data that cannot fit in the computer memory, selecting optimum sliding window size, recognizing sessions, and searching/extracting domain knowledge. Prediction challenges are memory limitation, extensive training/prediction time and low prediction accuracy. To address these challenges, I want to implement the web page prediction problem using MapReduce programming model of Hadoop

framework. Apache Hadoop is an open source software framework for storage and large scale processing of data sets. MapReduce programming model of the Hadoop framework has the ability to rapidly process large amount of data in parallel.

2. Literature Survey

2.1 Applications

Web page prediction problems can be generalized and applied in many essential industrial applications and in personalization applications [1]. Industrial applications include applications such as search engines, caching systems, recommendation systems and wireless applications**.** Here, users are categorized based on their interests and tastes. The previously visited categorized Web pages are used to capture interests and tastes of users.

2.2 Related Work

The basic concept of the Markov model [5], [6] is to predict the next action, depending on the result of previous actions. In Web prediction, the next action corresponds to predicting the next page to be visited. The previous actions correspond to the previous pages that have already been visited. The previous actions correspond to the previous pages that have already been visited. In web prediction, the *K*th-order Markov model is the probability that a user will visit the *k*th page, provided that he/she has visited $k - 1$ pages.

In all-*K*th Markov model [5], [6], we generate all orders of Markov models and utilize them collectively in prediction. Compare to Markov model, the all-Kth-order Markov model achieves better prediction and it only fails when all orders of the basic Markov models fail to predict.

The basic idea in the modified Markov model [5] is to consider a set of pages in building the prediction model to reduce its size. So it doing this by reducing the number of path in the model so that it can fit in the memory and predict faster. The main objective is that a task on the Web can be done using different paths regardless of the ordering that users choose. In addition, we reduce the size of prediction model by discarding sessions that have repeated pages.

ARM [5],[8] is a data mining technique that has been applied successfully to discover related transactions. In ARM, relationships among item sets are discovered based on their concurrence in the transactions. Specifically, ARM focuses on associations among frequent item sets. In WPP, prediction is conducted according to the association rules that satisfy certain support and confidence as follows.

Mobasher et al. [9] use the ARM technique in WPP and propose the frequent item set graph to match an active user session with frequent item sets and predict the next page that user is likely to visit.

A Web Based Recommendation using Association rule and clustering was proposed by Singhal and Panday[8].

k-Nearest Neighbors algorithm [18](or *k*-NN for short) is a non-parametric method used for classification and regression. Anitha et al.[10] use integrated approach like upper approximation based rough set clustering using k nearest neighbors, dynamic support pruned all k-th order Markov model and all k-th order association rule mining by dynamic frequent (k+1) item set generation using Apriori.

Nasraoui and Petenes [11] proposed an approach based on fuzzy inference to provide Web Recommendations. The proposed approach is based on rule that are automatically derived from prediscovered user profiles.

Jalali et al. [12] propose a new classification model for online predicting user's future movements which is based on Longest common Subsequence concept. In his research work, they divides architecture in two main phase: offline phase and online phase.

Agarwal et al. [2] propose an efficient weighted algorithm for web information retrieval system. The proposed system will use frequency of a page, time spent on a page and click history of a page to assign a quantitative weight to each page for a user.

Awad et al. [13] proposed hybrid approach using artificial neural network, and the All-*K*th Markov model, to resolve prediction using Dempster's rule. Such fusion overcomes the inability of the Markov model in predicting beyond the training data, as well as boosts the accuracy of ANN, particularly, when dealing with a large number of classes. In their proposed work, they used backpropagation algorithm for multilayer neural network learning. Pruthvi R [1] proposed web user's browsing behavior prediction based on neural network and implemented in

MapReduce.

Awad et al. [6] proposed hybrid approach using Support Vector Machines, and the All-*K*th Markov model, to resolve prediction using Dempster's rule. They apply feature extraction to increase the power of discrimination of SVM. In addition, during prediction we employ domain knowledge to reduce the number of classifiers for the improvement of accuracy and the reduction of prediction time.

3. Background Study

3.1 Big Data

Big data is a term for massive data sets, a large amount of data available in complex structured or no structure form. These vast amounts of data are generated by social media and networks, scientific instruments, mobile devices, sensor technology an networks. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics [3].

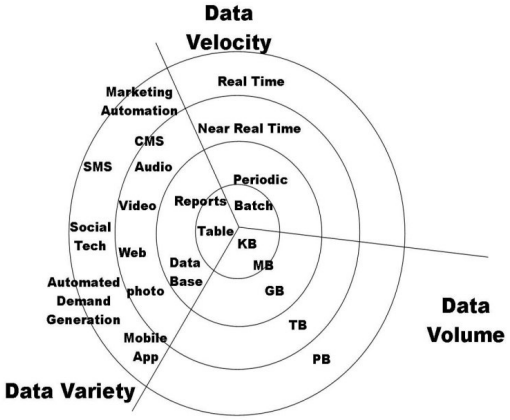

Figure 3.1: The three V's of Big data [3]

3.2 Apache Hadoop Framework

Apache Hadoop [14] is an open-source software framework for distributed storage and distributed processing of Big Data on clusters of commodity hardware. The Apache™ Hadoop® [15] project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop framework includes following modules:

- Hadoop Common: It having utilities that support the other hadoop module.
- Hadoop distributed File System (HDFS): a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce: a programming model for large scale data processing.

3.2.1 Hadoop Distributed File System

Hadoop Distributed File System is extended version of the Google's Google File System (GFS). The work of HDFS is responsible for storing the data on cluster of machines. The Hadoop runtime system coupled with HDFS manages the details of parallelism and concurrency to provide ease of parallel programming with reinforced reliability. In hadoop cluster, a master node controls a group of slave nodes on which the Map and Reduce functions run in parallel. The master node assigns a task to a slave node that has any empty task slot. There is single master node and multiple slave nodes possible in HDFS. Master node contains meta information of file. HDFS divide the data into 64MB block and divide among the nodes in cluster [3].

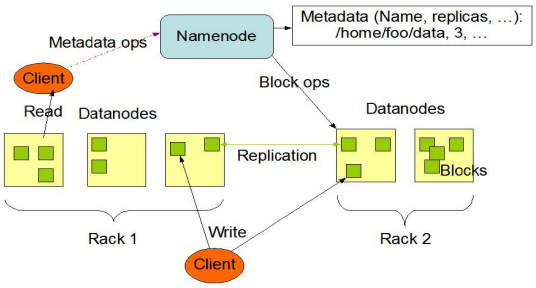

Figure 3.2 : Architecture of HDFS [16]

Differnet Nodes and daemons run on hadoop cluster are:

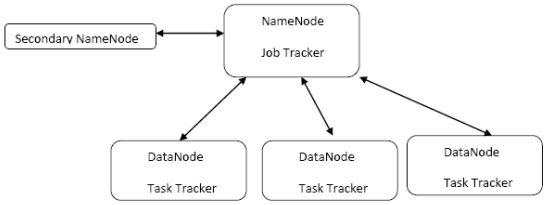

Figure 3.3 : Hadoop daemons and node

3.2.2 MapReduce

MapReduce [4], [19] is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key.

On top of HDFS, Hadoop MapReduce is the execution framework for MapReduce applications. MapReduce consists of a single master node called JobTracker, and worker nodes called TaskTrackers. Note that MapReduce TaskTrackers run on the same set of nodes that HDFS DataNodes run on [4].

3.3 Support Vector Machines

A Support Vector Machine (SVM) [7], [17] performs *classification* by constructing an *N*-dimensional hyperplane that optimally separates the data into two categories. In the reference of SVM literature, a predictor variable is called an *attribute*, and a transformed attribute that is used to define the hyperplane is called a *feature*. The task of choosing the most suitable representation is known as *feature selection*. A set of features that describes one case (i.e., a row of predictor values) is called a *vector*.

So the goal of SVM modeling is to find the optimal hyperplane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyperplane are the *support vectors*. An SVM analysis finds the line (or, in general, hyperplane) that is oriented so that the margin between the support vectors is maximized.

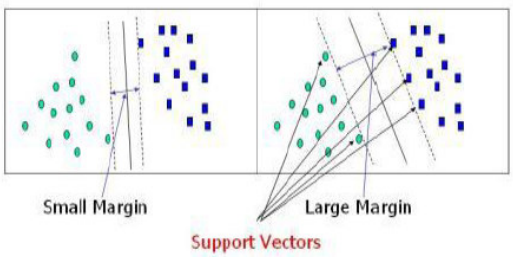

Figure:3.4 Support Vector

3.3.1 Multiclass SVM

In case of multiclass data sets, there are two primary schemes, namely :

- One-VS-one
- One-VS-all

Both are extensions to the binary classification upon which SVM is based.

The one-VS-one approach creates a classifier for each pair of classes. The total number of classifiers computed is $n(n-1)/2$, where n is the number of classes in the data set. A new instance x belongs to the class upon which most classifiers agree, i.e., majority voting.

One-VS-all creates a classifier for each class in the data set against the rest of the classes.

3.3.2 Why SVMs ??

- SVMs classification shows greater accuracy in predicting seen as well as unseen data as compare to Markov model
- Diversity of the kernel tricks for different problems.
- Stable with the changes on data
- High accuracy

4. Design of Experiments

For classification using support vector machine, WEKA Tool is available. Open-Source Software Tool is a collection of machine learning algorithms implemented in Java developed at the University of Waikato, New Zealand.

For the experiments we used, WEKA version 3.6, Operating System: 64-bit Windows 7 Home Premium with Intel Core i5 CPU @ 2.30 GHz and 3 GB of RAM.

Also we were took the preprocessed dataset in ARFF format from the library of university of British Columbia(UBC).

Dataset name: Amazon
Record: 1050, 1065
Class : 50
We used LIBSVM library in weka for the experiment purpose and set the parameter like SVM type , kernel type etc.

4.1 Experiment And Result

No of Instance & Size of file	Training Time (sec)	Accuracy
1050 (20.3 MB)	55.36	37.5556 %
1065 (20.6 MB)	58.78	37.5556 %

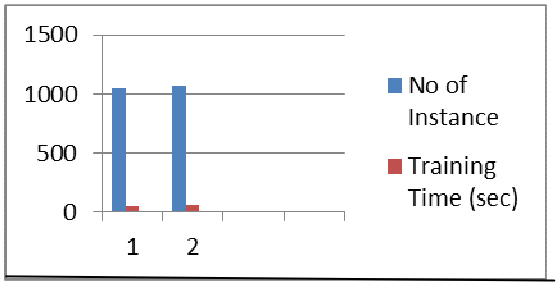

Figure:4.1 Results of small experiments

Awad et al. [6] perform the experiments and saw that for 5430 different web page ,they need to create 5430 classifier ,so total time of training for SVM is 26.3 h

4.2 Problem Gap

From the above result we noted, when the size of training data increase it will increase the training time

proportionally and having the lower prediction accuracy. At some point weka will not work because of computation power and memory limitation of system itself. So we have to come up with new approach that parallelize the computation and improve the accuracy and lower the training time when source of input data is too large.

5. Conclusion

Web page prediction is important since the prediction of the user's browsing behavior reduces the browsing access time and avoids the visit of unnecessary pages, to ease network traffic. According to our observations, Support vector machine and many other machine learning algorithms do not fit when source of input data is too large. Now a day data is generated day by day in drastic manner. The entire present classification algorithm is having extensive training and prediction time and lower prediction accuracy when the Big data is source of input. So we have to proposed parallel support vector machines for web page prediction based on MapReduce programming model and runs on Hadoop framework. It removes scalability problem of present SVM algorithm And currently is an active research area for web usage mining from bigdata. In near future we will implement the proposed solution in hadoop framework by considering large datasets. And comparing the result with existing algorithm and different node of hadoop framework.

Reference

Papers:

[1] Pruthvi R "Web-Users' Browsing Behavior Prediction by Implementing Neural Network in MapReduce", IJAFRC vol. 1, issue 5, May 2014.

[2] Agarwal, Rohit; Arya, K.V.; Shekhar, Shashi; Kumar, Rakesh "An Efficient Weighted Algorithm for Web Information Retrieval System" IEEE 2011 International Conference on Computational Intelligence and Communication Networks (CICN) , pp. 126-131, 2011.

[3] Dagli Mikin K., and Brijesh B. Mehta, "Big Data and Hadoop: A Review", IJARES Volume2, Issue2, February 2014, pages 192-196.

[4] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1):107–113, 2008.

[5] Mamoun A. Awad and Issa Khalil, "Prediction of User's Web-Browsing Behavior: Application of Markov Model", IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, vol. 42, no. 4, pp. 1131-1142, August 2012.

[6] M. Awad, L. Khan, and B. Thuraisingham, "Predicting WWW surfing using multiple evidence combination," *VLDB J.*, vol. 17, no. 3, pp. 401– 417, May 2008.

[7] Kiran M, Amresh Kumar, Saikat Mukkherjee, and Ravi Prakash G, "Verification and Validation of MapReduce Program Model for Parallel Support Vector Machine Algorithm on Hadoop Cluster", IJCSI, Vol. 10 Issue 3, No. 1, May 2013.

[8] Vidhu Singhal, Gopal Pandey, "A Web Based Recommendation Using Association Rule and Clustering", IJCCER, Vol. 1, Issue 1, May 2013

[9] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Effective personalization based on association rule discovery from Web usage data," in *Proc. ACM Workshop WIDM*, Atlanta, GA, Nov. 2001.

[10] A.Anitha, N.Krishnan, "A Web Usage Mining based Recommendation Model for Learning Management Systems", 978-1-4244-5967-4, IEEE, 2010

[11] O. Nasraoui and C. Petenes, "Combining Web usage mining and fuzzy inference for Website personalization," in *Proc. WebKDD*, 2003, pp. 37–46.

[12] Jalali, Mehrdad; Mustapha, Norwati; Mamat, Ali; Sulaiman, Md. Nasir B, "A new classification model for online predicting users' future movements", International Symposium on Information Technology,IEEE ,pp. 1-7,2008.

[13] M. Awad and L. Khan, "Web navigation prediction using multiple evidence combination and domain knowledge," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 37, no. 6, pp. 1054–1062, Nov. 2007.

Website

[14]Apache Hadoop, December 2014: http://en.wikipedia.org/wiki/Apache_Hadoop

[15] Welcome to Apache Hadoop, December 2014: http://hadoop.apache.org/

[16] HDFS Architecture Guide, April 2013: http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[17] Support Vector Machines, December 2014 : http://en.wikipedia.org/wiki/Support_vector_machine

[18] *k*-nearest neighbors algorithm, November 2014 : http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Books:

[19] Tom White, Hadoop the definitive guide Yahoo Press, second edition, 2011

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar